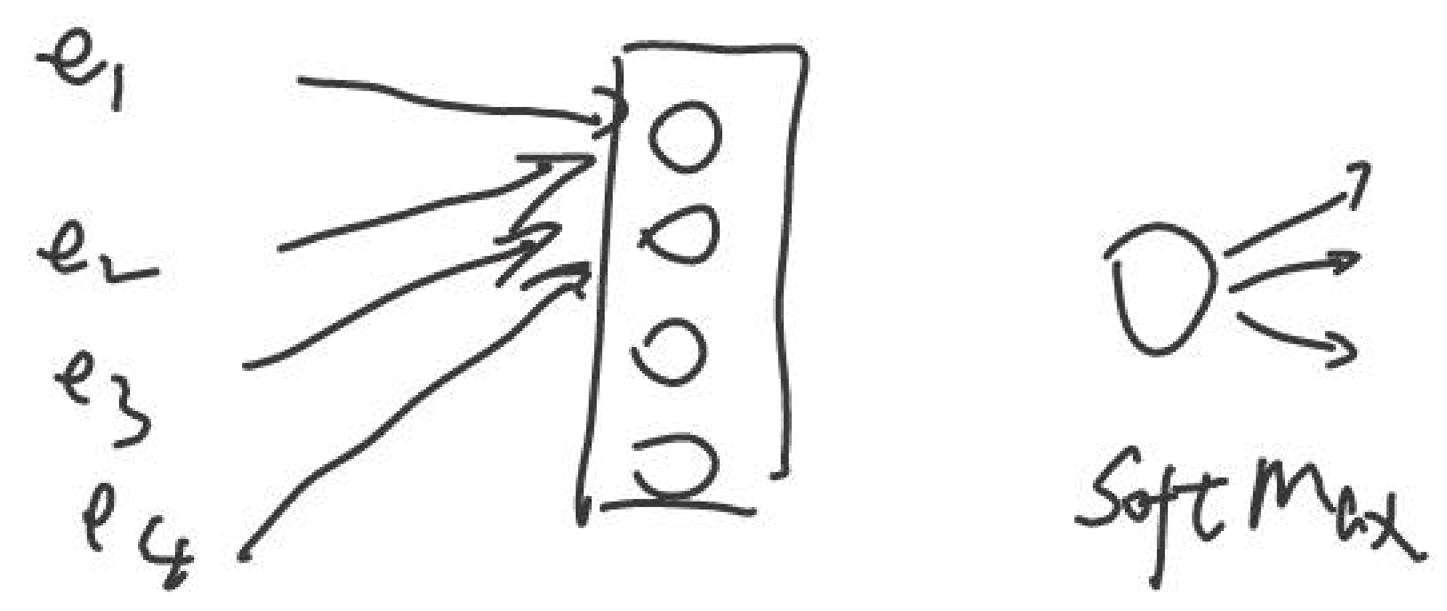
1 15

Word Embeddings.

$\underbrace{\text{a glass of orange}}$ ______ ?



10000×1.

windows size = 4 vocabulary size = 10,000.

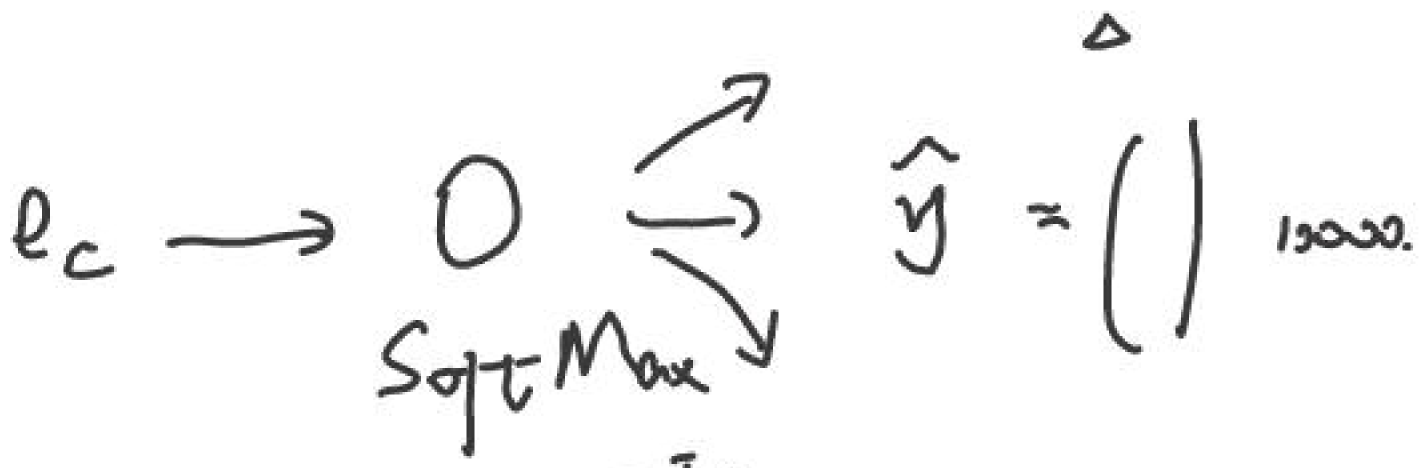
训练 E, w, b.

2003, neaural probabilistic language model.

$\boxed{\text{Skip}}$ - Gram.

给定中心词 预测上下文. "context" "target"!

a glass of orange juice.



$$p(t|c) = \frac{e^{\theta_t^T e_c}}{\sum_j e^{\theta_j^T e_c}}.$$

Problems: 计算成本高. ⇒ Hierachi: SoftMax

对c的采样也用启发式, 树型分类器.

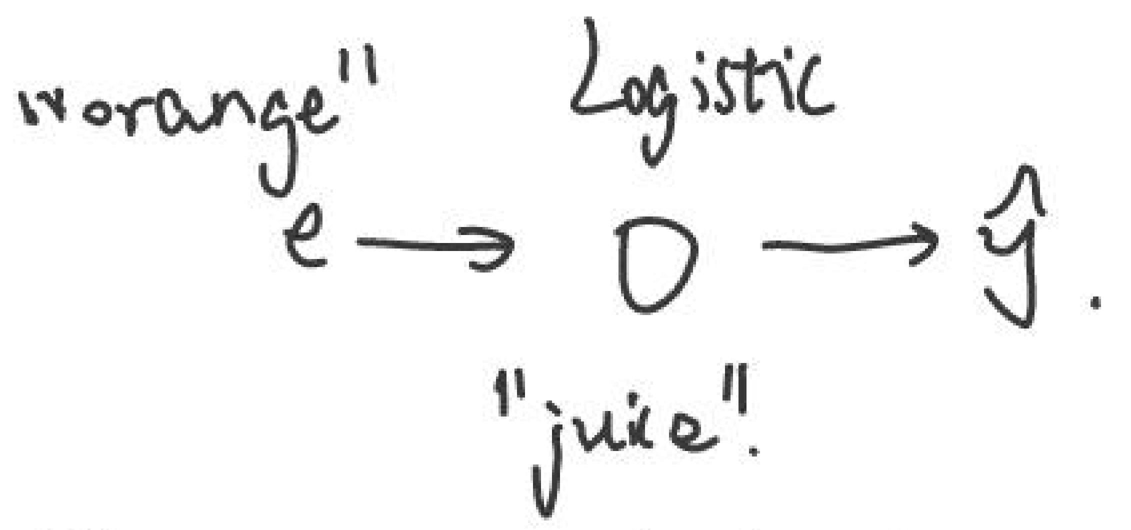
避免 "the、of, a, an"!

1 15

Negative Sampling.

正样本：负样本 = 1:k，随机采样.

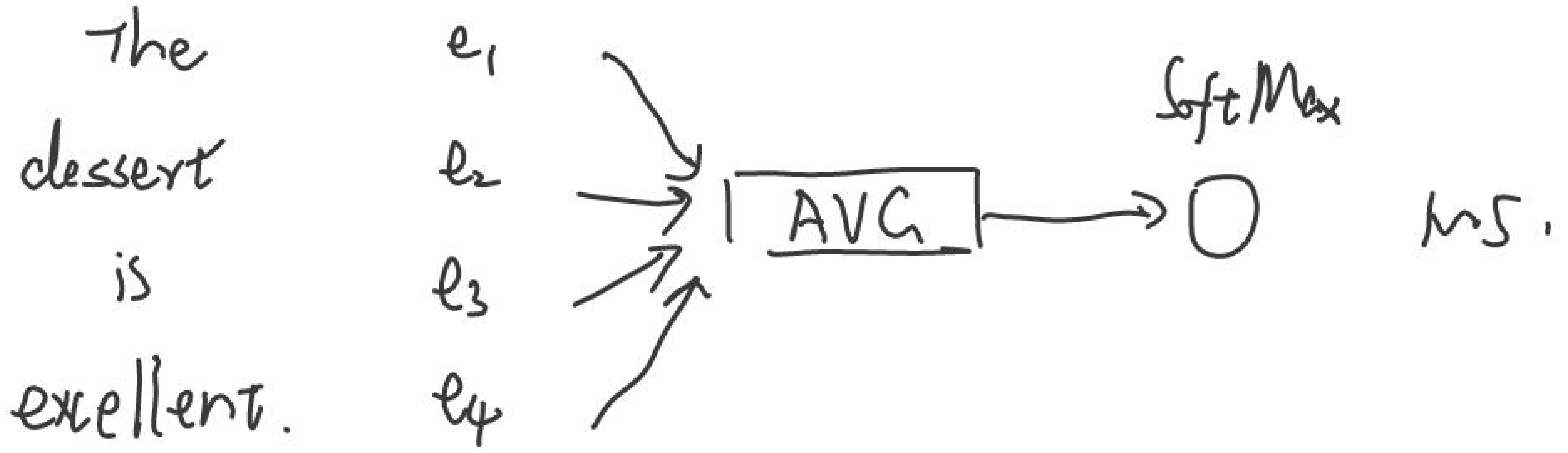
$P(y=1 \mid t, c) = \sigma(\theta_t^T \cdot e_c)$.



采样策略：在均匀分布与频率分布之间最"合适".

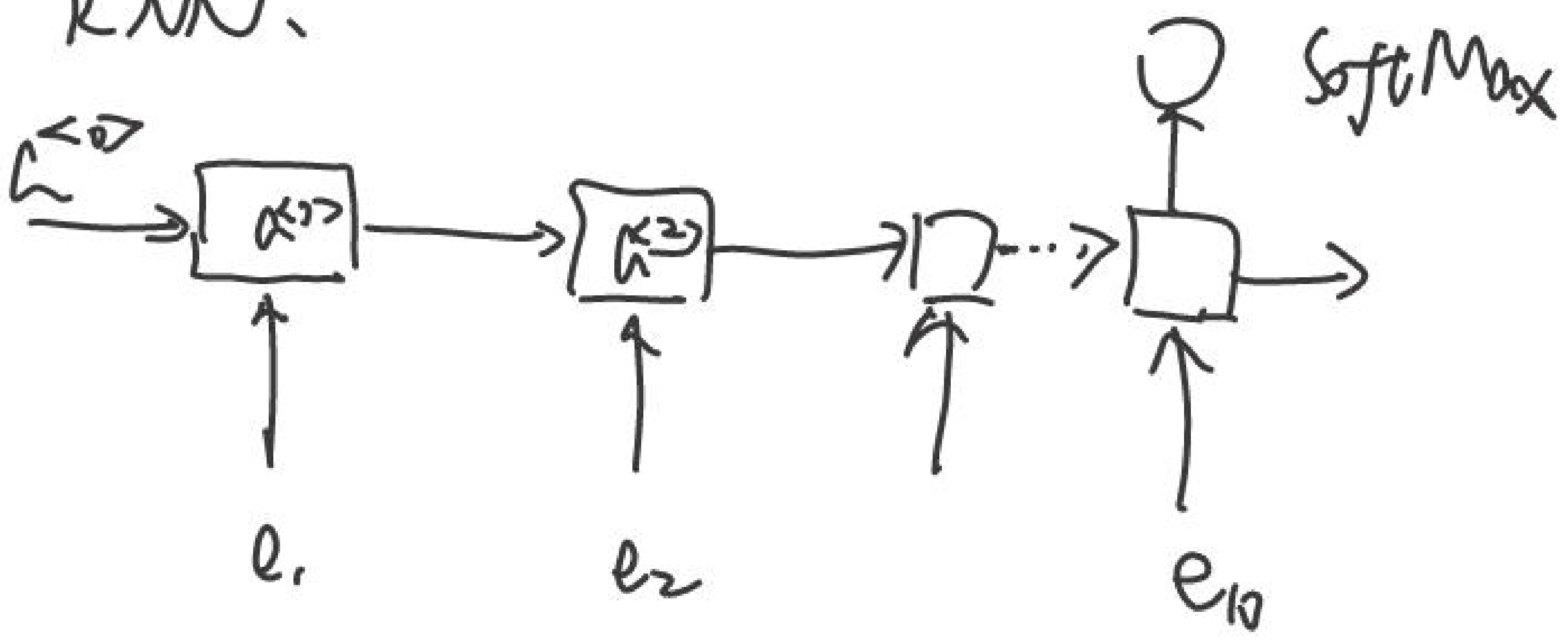
Sentiment Classification.

1):



受词出现次数的影响.

2): RNN、



1 15

Glove Word Embeddings.

X_{ij}: i 出现在 j的上下文中的次数.

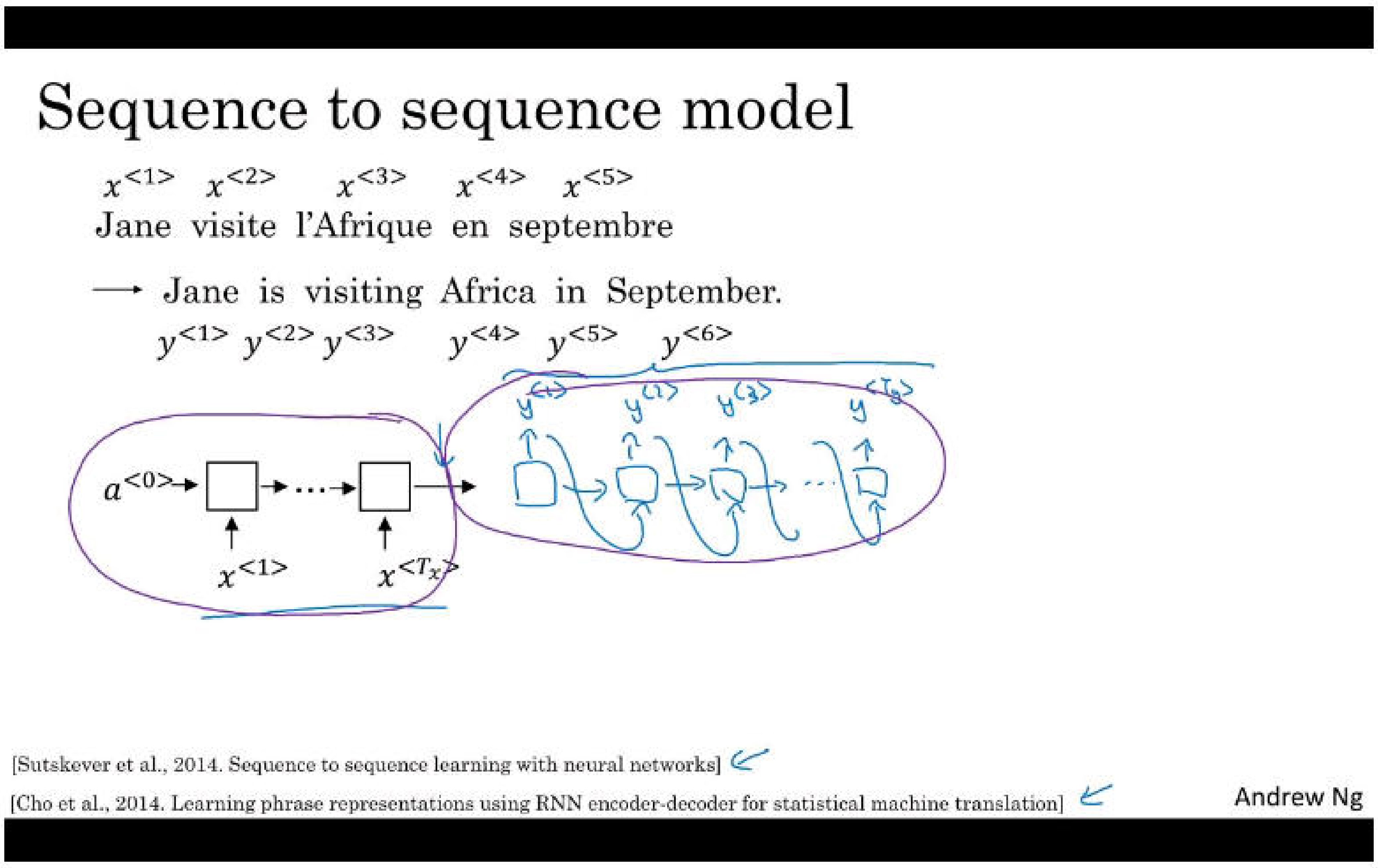
minimize $\sum_i \sum_j \underset{\downarrow}{f(X_{ij})}(\theta_i^T e_j^{\uparrow} + b_i + b_j - \lg X_{ij}^{\uparrow})^2$

加权: $X_{ij}=0$; "the, a, an".

θ_i与e_j是完全对称的, $e_w = \frac{e_w+\theta_w}{2}$.

1 1b. Sequence to Sequence.

Basic Models.



Sequence to sequence model

$x^{<1>}$ $x^{<2>}$ $x^{<3>}$ $x^{<4>}$ $x^{<5>}$
Jane visite l'Afrique en septembre

→ Jane is visiting Africa in September.
$y^{<1>}$ $y^{<2>}$ $y^{<3>}$ $y^{<4>}$ $y^{<5>}$ $y^{<6>}$

[Sutskever et al., 2014. Sequence to sequence learning with neural networks]
[Cho et al., 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation]

Andrew Ng

Encoder - Decoder 结构 同样被用在了 Image Caption

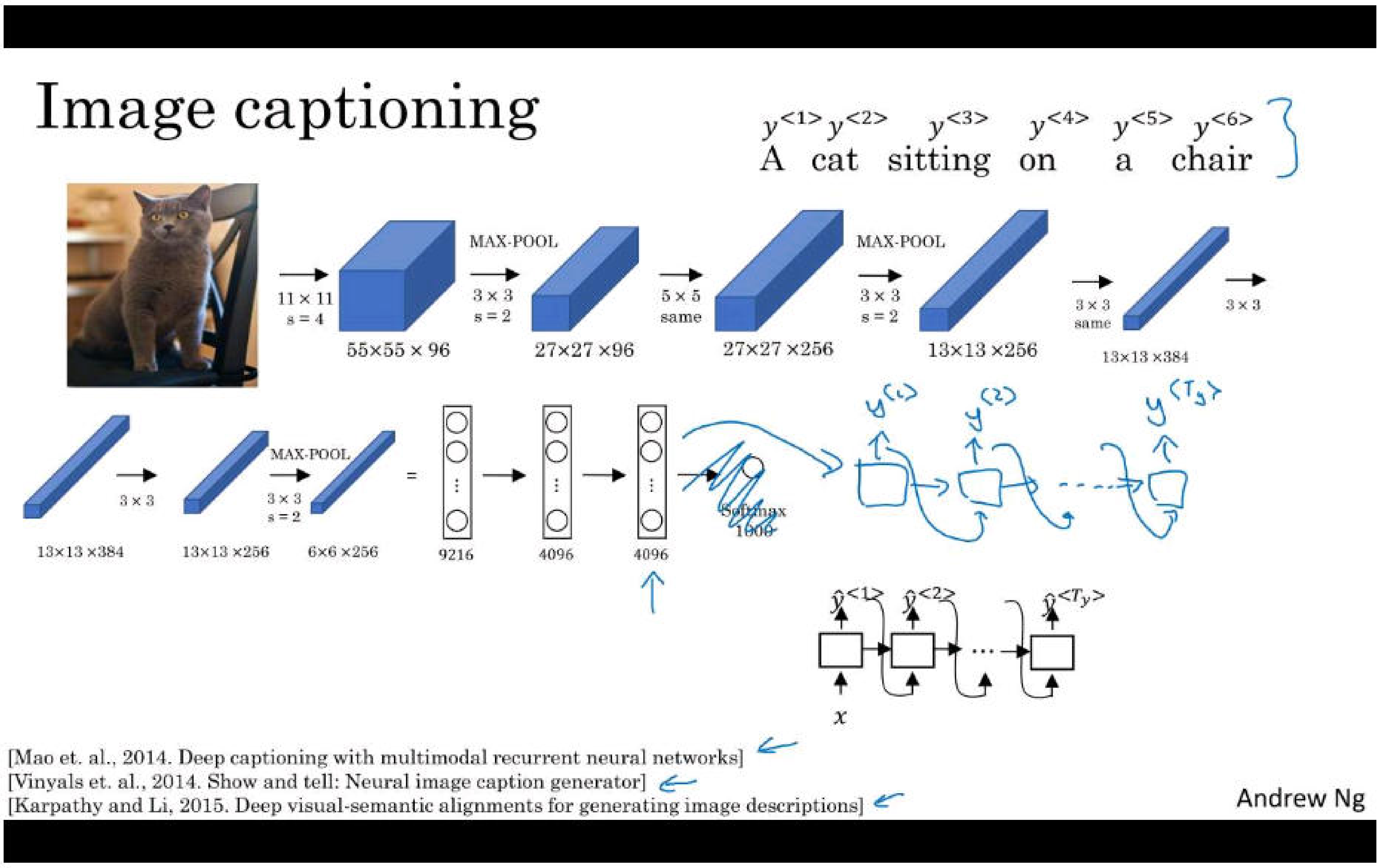


Image captioning

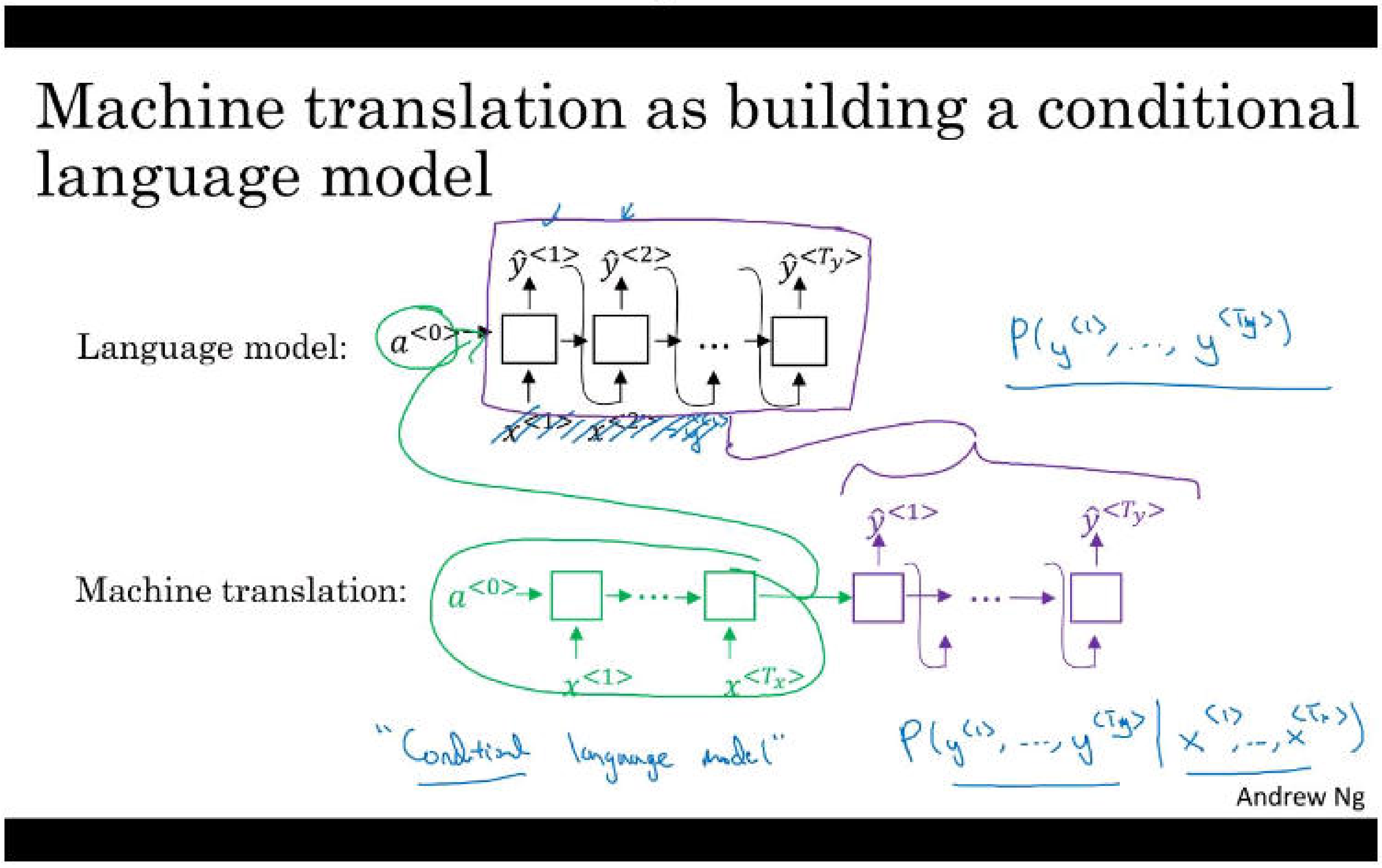
$y^{<1>}$ $y^{<2>}$ $y^{<3>}$ $y^{<4>}$ $y^{<5>}$ $y^{<6>}$
A cat sitting on a chair

[Mao et. al., 2014. Deep captioning with multimodal recurrent neural networks]
[Vinyals et. al., 2014. Show and tell: Neural image caption generator]
[Karpathy and Li, 2015. Deep visual-semantic alignments for generating image descriptions]

Andrew Ng

1 16

Machine translation.



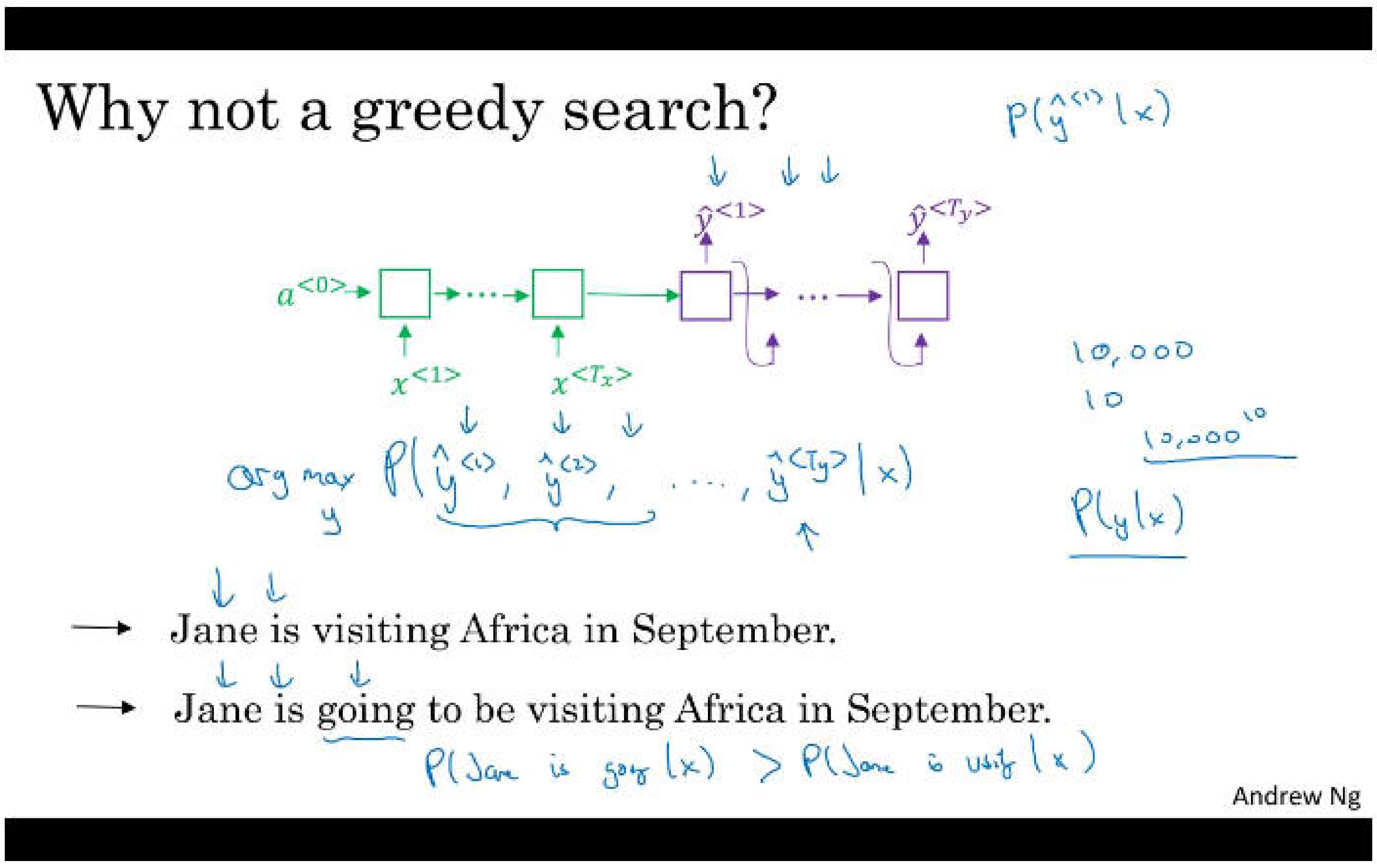
Machine translation as building a conditional language model

Language model:
$a^{<0>}$
$\hat{y}^{<1>}$ $\hat{y}^{<2>}$ … $\hat{y}^{<T_y>}$
$x^{<1>}$ $x^{<2>}$
$P(y^{<1>},\dots,y^{<T_y>})$
Machine translation:
$a^{<0>}$
$x^{<1>}$ $x^{<T_x>}$
$\hat{y}^{<1>}$ $\hat{y}^{<T_y>}$

"Conditional language model" $P(y^{<1>},\dots,y^{<T_y>} \mid x^{<1>},\dots,x^{<T_x>})$

Andrew Ng

条件语言模型 源语句是条件.



Why not a greedy search?

$\underset{y}{\arg\max}\ P(\underbrace{\hat{y}^{<1>}, \hat{y}^{<2>},}\ \dots, \hat{y}^{<T_y>} | x)$

→ Jane is visiting Africa in September.

→ Jane is going to be visiting Africa in September.

P(Jane is going | x) > P(Jane is visiting | x)

Andrew Ng

输出是随机采样的. $\underset{y}{\arg\max}\ P(y|x)$

不能用Greedy, 会非 全局最优.

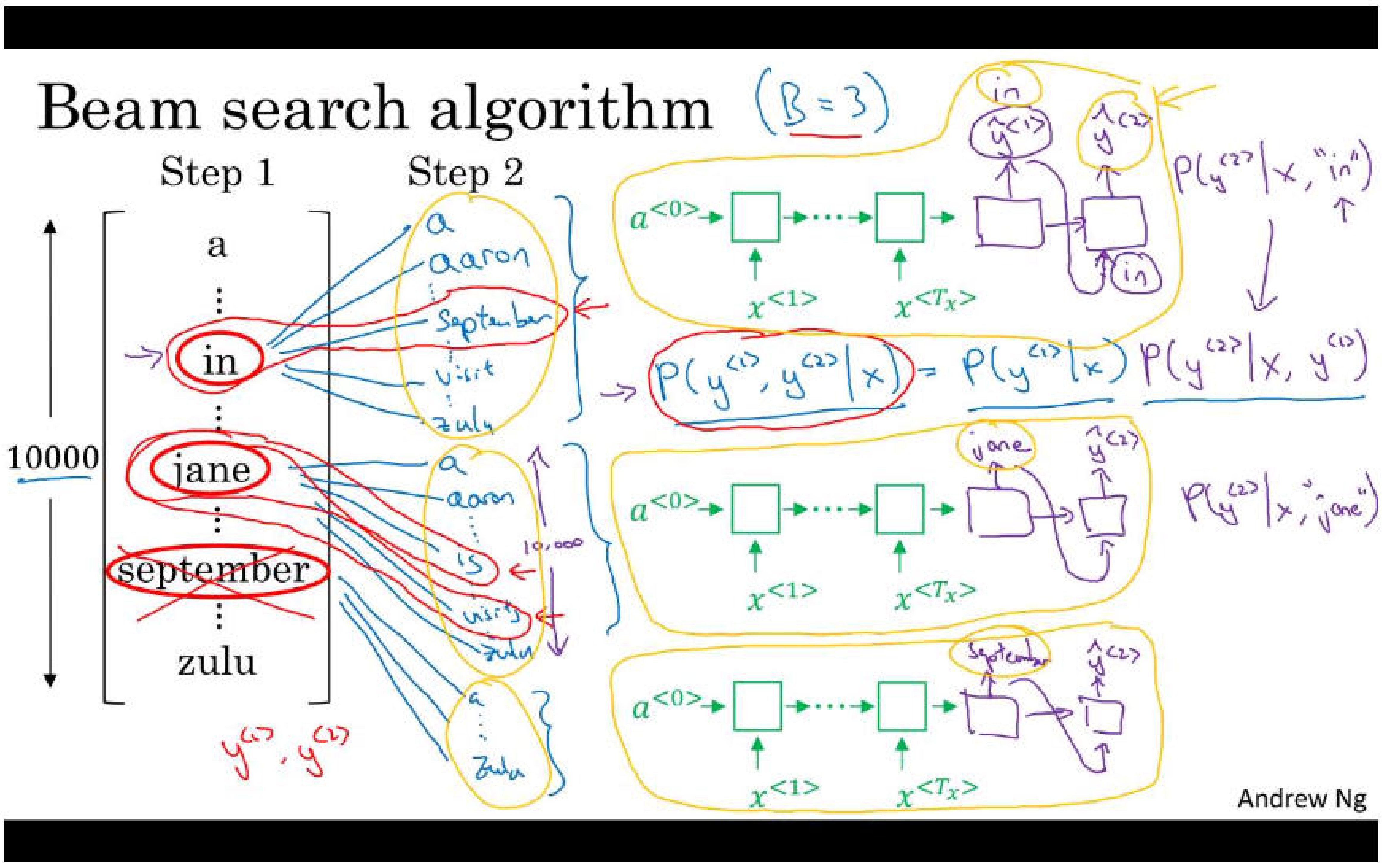
↓

Beam Search.

1 16

Beam Search.

每次保留 B = Beam width 个候选词.



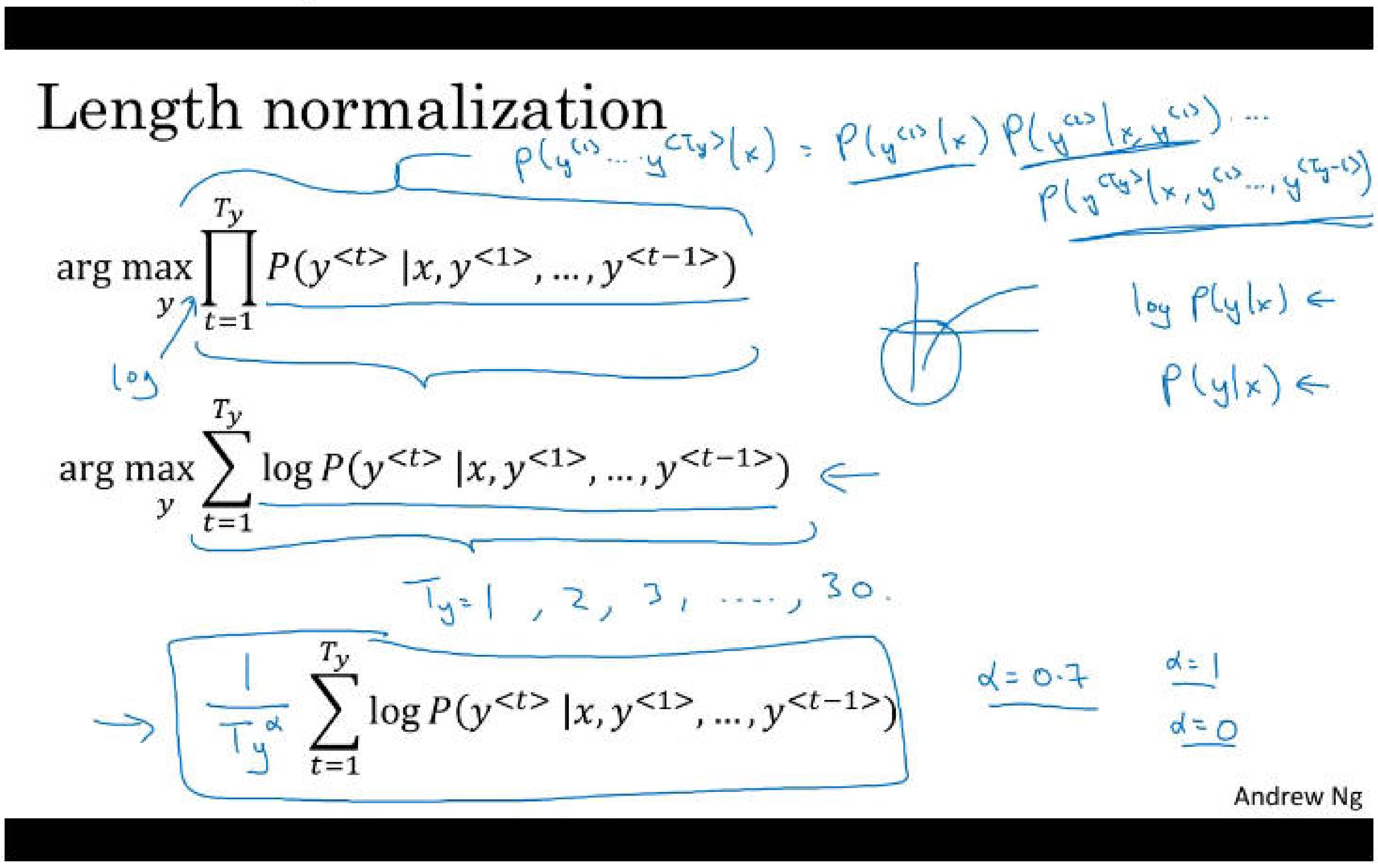
Beam search algorithm (B=3)

RNN 给出了 $P(y^{<1>}|x)$ $P(y^{<2>}|x, y^{<1>})$

机器翻译想获得的是 $\arg\max P(y^{<1>}, y^{<2>}|x)$

$P(y^{<1>}, y^{<2>}|x) = P(y^{<2>}|y^{<1>}, x) \cdot P(y^{<1>}|x)$.

每次保留 B 个.



Length normalization

$P(y^{<1>} \dots y^{<T_y>}|x) = P(y^{<1>}|x) P(y^{<2>}|x, y^{<1>}) \dots P(y^{<T_y>}|x, y^{<1>}, \dots, y^{<T_y-1>})$

$$\arg\max_y \prod_{t=1}^{T_y} P(y^{<t>}|x, y^{<1>}, \dots, y^{<t-1>})$$

log

$$\arg\max_y \sum_{t=1}^{T_y} \log P(y^{<t>}|x, y^{<1>}, \dots, y^{<t-1>})$$

$T_y = 1, 2, 3, \dots, 30.$

$$\frac{1}{T_y^{\alpha}} \sum_{t=1}^{T_y} \log P(y^{<t>}|x, y^{<1>}, \dots, y^{<t-1>})$$

$\log P(y|x)$ ←

$P(y|x)$ ←

$\alpha = 0.7$

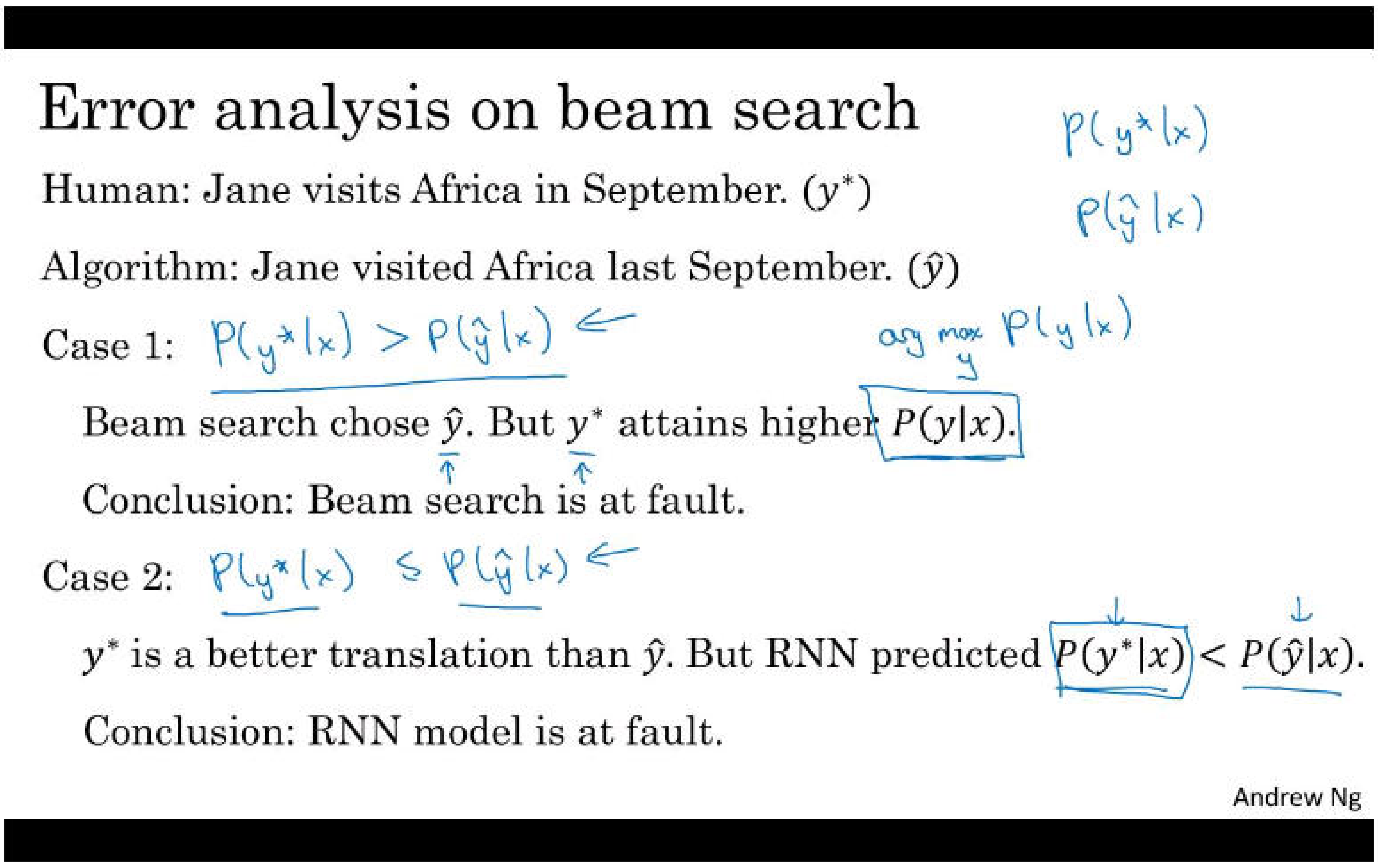
$\alpha = 1$

$\alpha = 0$

Andrew Ng

对 $\arg\max P(y|x)$ 正则 ⇒ $\arg\max \frac{1}{T_y^{\alpha}} \cdot \Sigma \log P(\ )$.

1 16



Error analysis on beam search

$P(y^*|x)$

$P(\hat{y}|x)$

Human: Jane visits Africa in September. (y^*)

Algorithm: Jane visited Africa last September. ($\hat{y}$)

Case 1: $\underline{P(y^*|x) > P(\hat{y}|x)}$ ← $\arg\max_{y} P(y|x)$

Beam search chose $\hat{y}$. But y^* attains higher $\boxed{P(y|x)}$.

Conclusion: Beam search is at fault.

Case 2: $P(y^*|x) \le P(\hat{y}|x)$ ←

y^* is a better translation than $\hat{y}$. But RNN predicted $\boxed{P(y^*|x)} < \underline{P(\hat{y}|x)}$.

Conclusion: RNN model is at fault.

Andrew Ng

对 Beam Search 出错的一些错误分析.

△ 机器翻译的评价指标：BLEU 值.

Intuition：与人类翻译越接近，质量越好.

Bilingual Evaluating Understudy (替补).

需要截断得分，否则 "the * 7".

n-grams：

$$P_n = \frac{\sum_{n\text{-grams}\in\hat{y}} \text{Count-Clip}(n\text{-grams})}{\sum_{n\text{-grams}\in\hat{y}} \text{Count}(n\text{-grams}}$$

$$BP = \begin{cases} 1 & \text{, if output\_length} > \text{ref\_length} \\ \exp(1 - \text{output\_length}/\text{ref\_length}) & \text{, otherwise.} \end{cases}$$

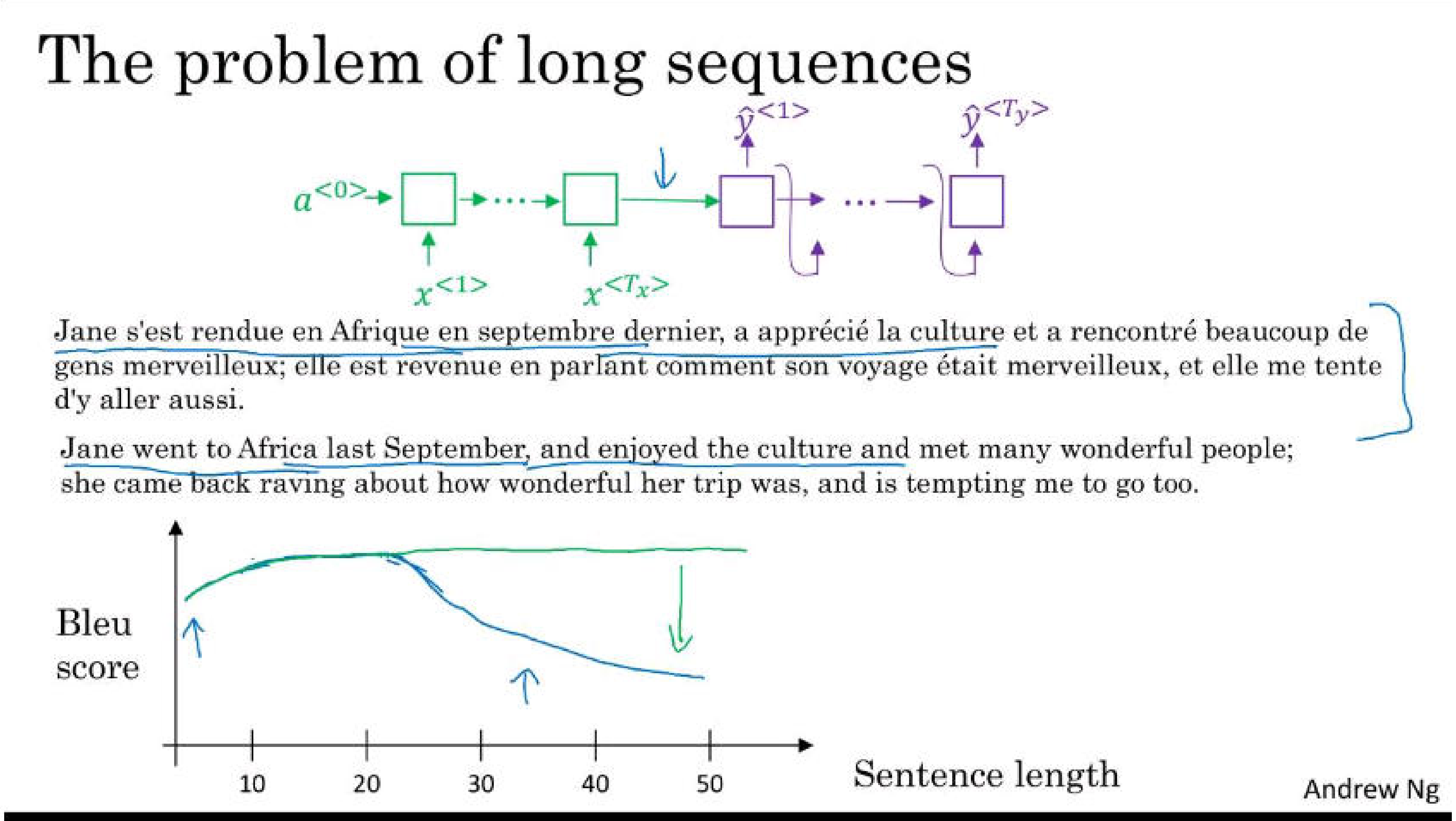
$$BLEU = BP \cdot \exp\left(\frac{1}{K}\sum_{n=1}^{K} P_K\right)$$

brevity penalty.

用 exp 做了加权.

1 16

Attention Mechanism.



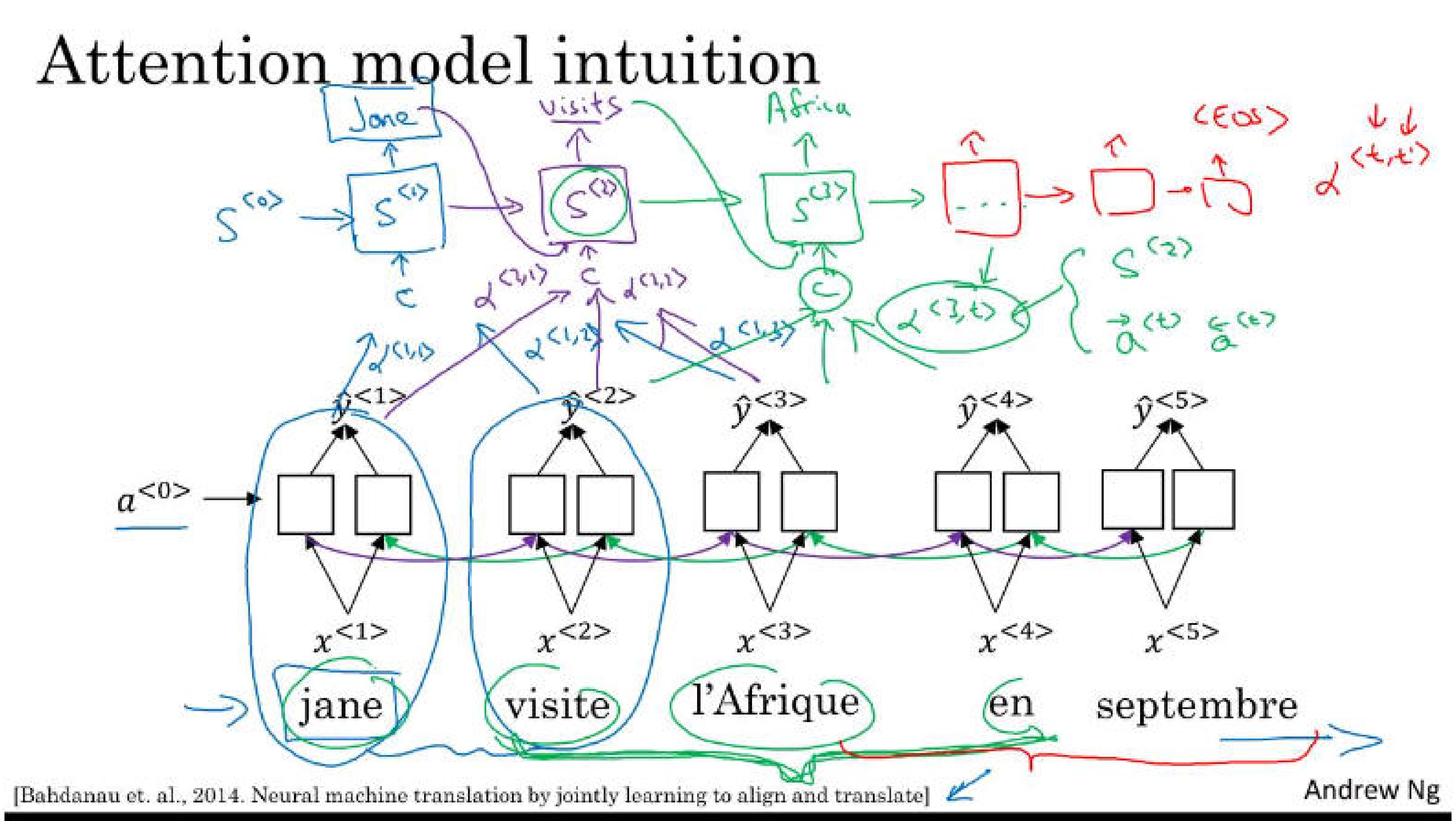
The problem of long sequences

Jane s'est rendue en Afrique en septembre dernier, a apprécié la culture et a rencontré beaucoup de gens merveilleux; elle est revenue en parlant comment son voyage était merveilleux, et elle me tente d'y aller aussi.

Jane went to Africa last September, and enjoyed the culture and met many wonderful people; she came back raving about how wonderful her trip was, and is tempting me to go too.

Andrew Ng

人做的是部分翻译，而RNN是整句翻译（Bad）.



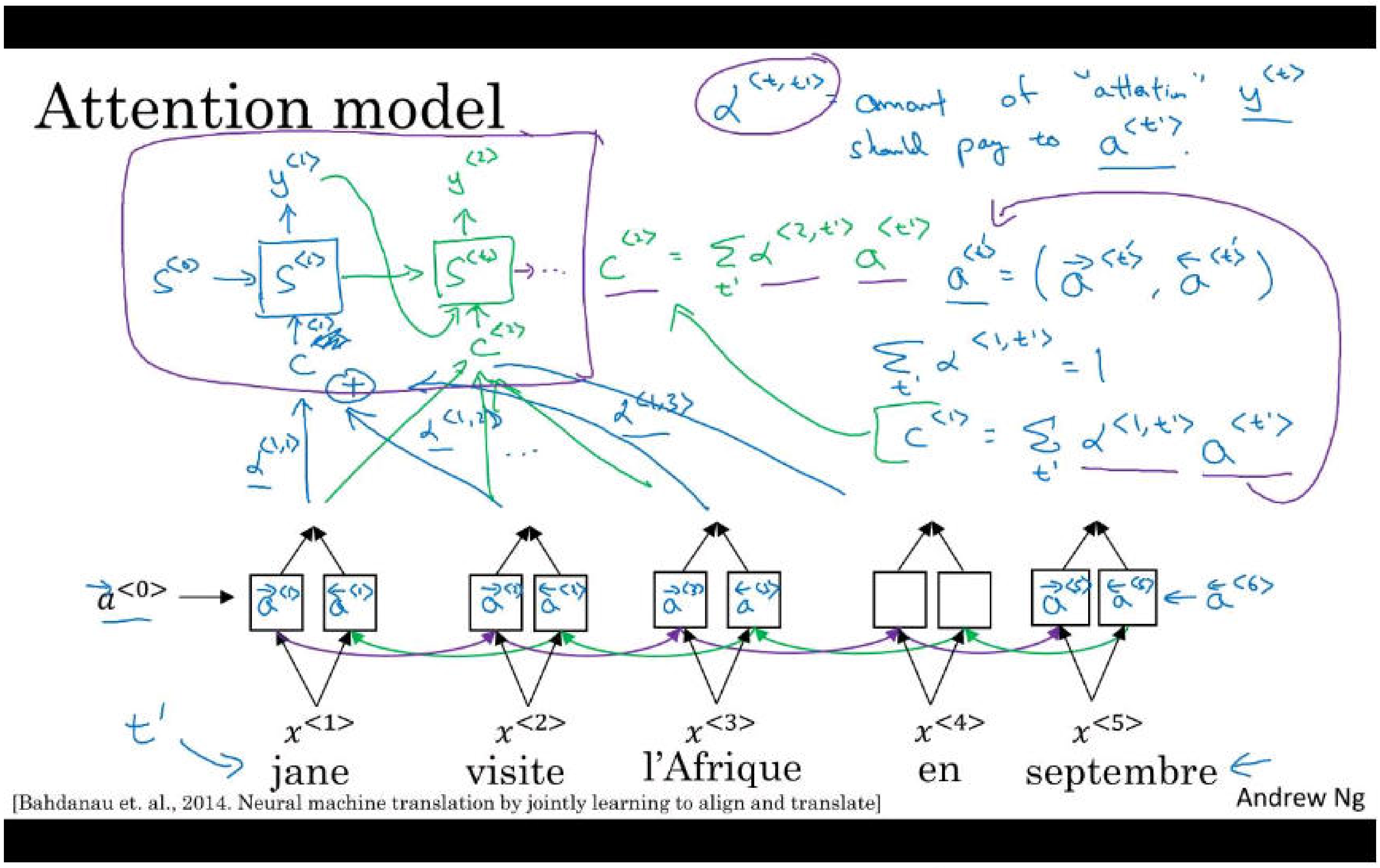
Attention model intuition

[Bahdanau et. al., 2014. Neural machine translation by jointly learning to align and translate]

Andrew Ng

Att 模拟部分对部分的翻译.

1 1b



Attention model

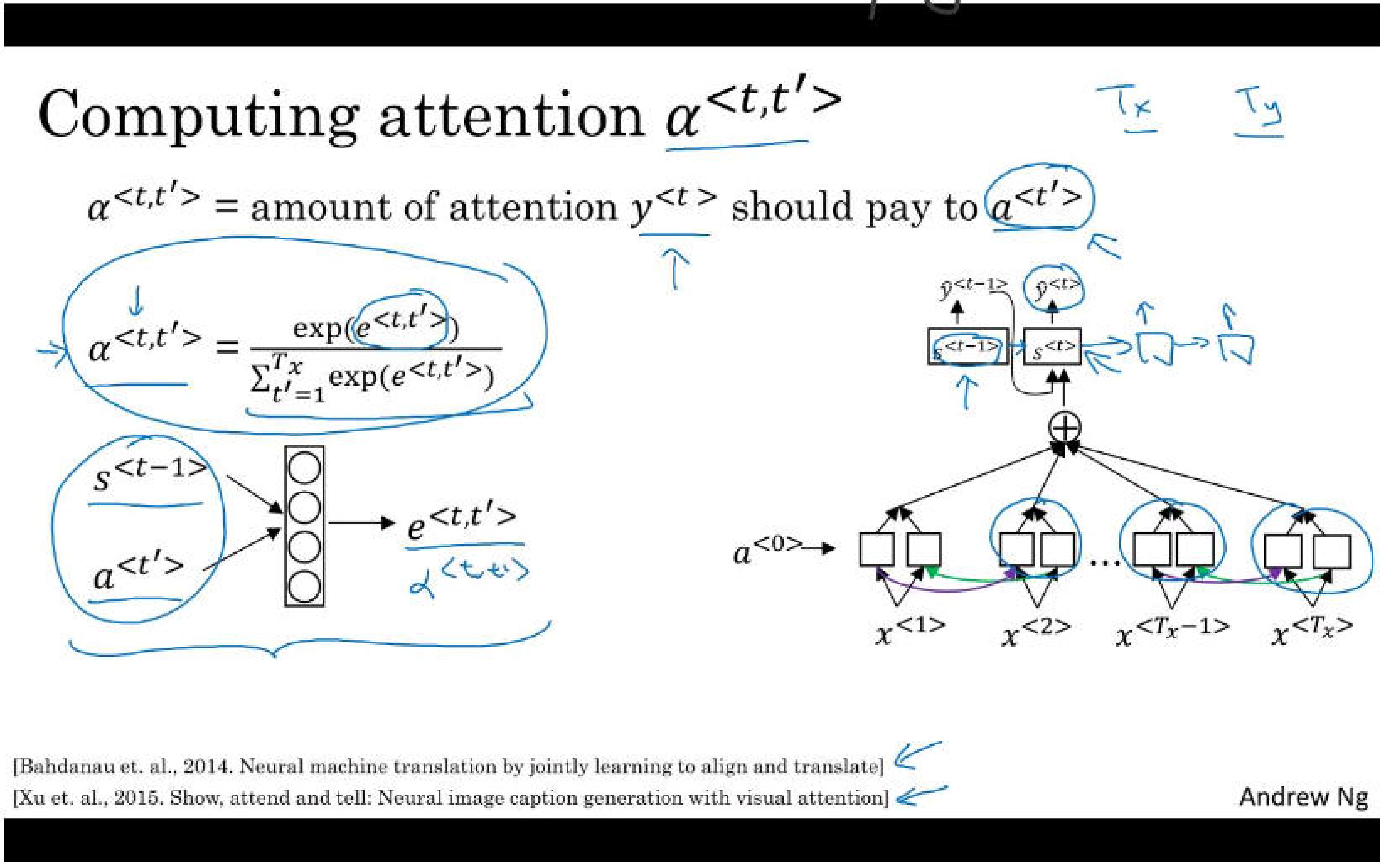
$\alpha^{<t,t'>}$ amount of "attention" $y^{<t>}$ should pay to $a^{<t'>}$.
$y^{<1>}$
$y^{<2>}$
$s^{<0>}$
$s^{<1>}$
$s^{<2>}$
$c^{<1>}$
$c^{<2>}$
$c^{<2>} = \sum_{t'} \alpha^{<2,t'>} a^{<t'>}$
$a^{<t'>} = (\overrightarrow{a}^{<t'>}, \overleftarrow{a}^{<t'>})$
$\sum_{t'} \alpha^{<1,t'>} = 1$
$c^{<1>} = \sum_{t'} \alpha^{<1,t'>} a^{<t'>}$
$\alpha^{<1,1>}$
$\alpha^{<1,2>}$
$\alpha^{<1,3>}$
$a^{<0>}$
$a^{<6>}$
t'
$x^{<1>}$
$x^{<2>}$
$x^{<3>}$
$x^{<4>}$
$x^{<5>}$
jane
visite
l'Afrique
en
septembre

[Bahdanau et. al., 2014. Neural machine translation by jointly learning to align and translate]

Andrew Ng

$\sum_{t'} \alpha^{<t,t'>} = 1$. $c^{<1>} = \sum_{t'} \alpha^{<1,t'>} a^{<t'>}$. C: Context.

$\alpha^{<t,t'>}$: attention $y^{<t>}$ pay to $a^{<t'>}$.



Computing attention $\alpha^{<t,t'>}$

$\alpha^{<t,t'>}$ = amount of attention $y^{<t>}$ should pay to $a^{<t'>}$

$$\alpha^{<t,t'>} = \frac{\exp(e^{<t,t'>})}{\sum_{t'=1}^{T_x} \exp(e^{<t,t'>})}$$

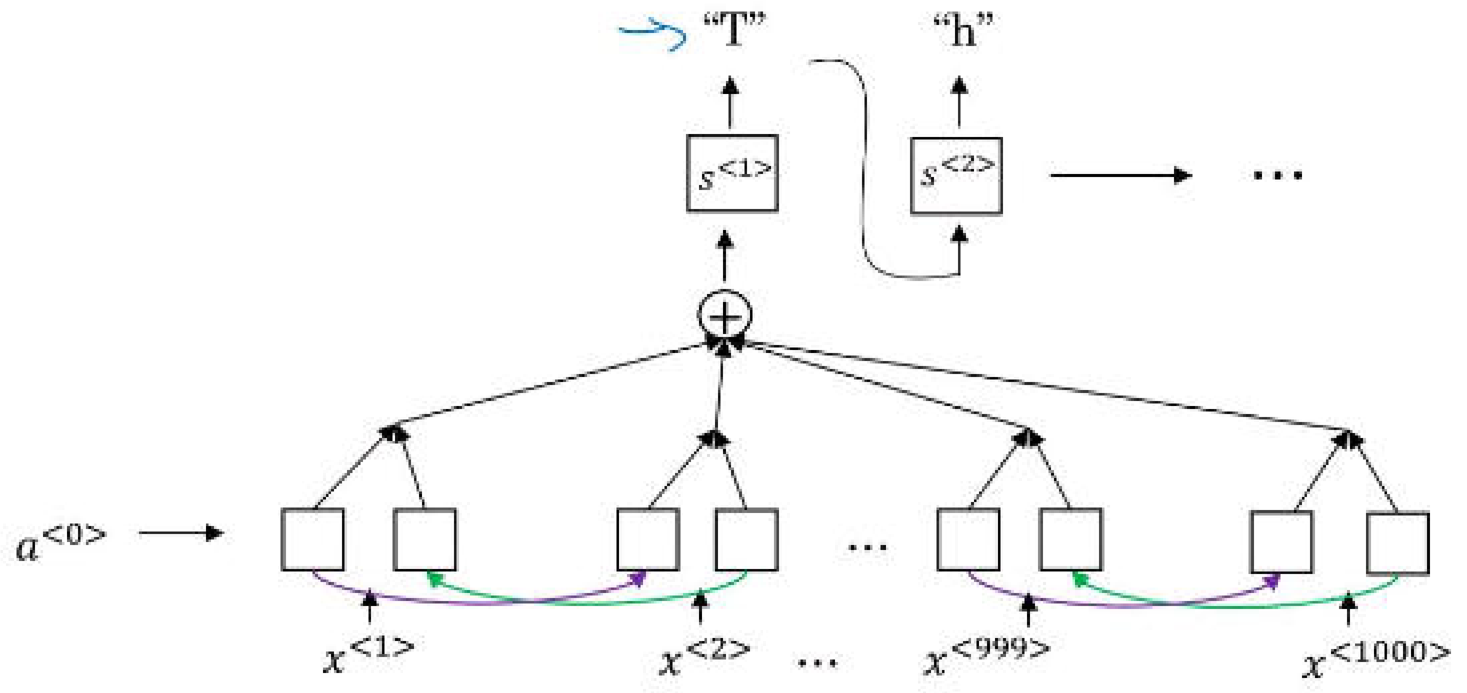
[Bahdanau et. al., 2014. Neural machine translation by jointly learning to align and translate]
[Xu et. al., 2015. Show, attend and tell: Neural image caption generation with visual attention]

Andrew Ng

Att 会与 $s^{<t-1>}$, $a^{<t'>}$ 相关. 但不知道. 构造小型NN.

1 16. Speech Recognition.

Attention model for speech recognition



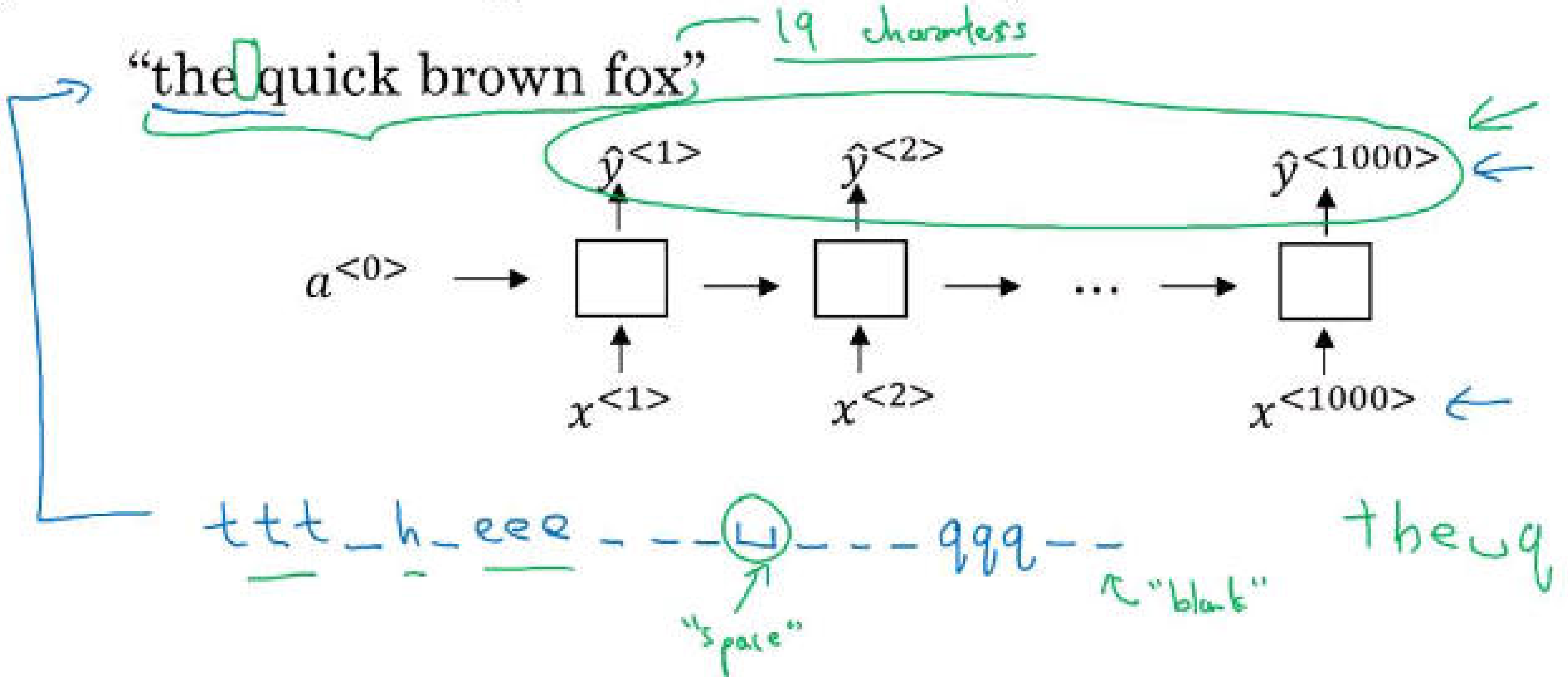
Andrew Ng

本质上也是一种翻译，也可以应用 Att.

CTC cost for speech recognition

(Connectionist temporal classification)

“the quick brown fox”



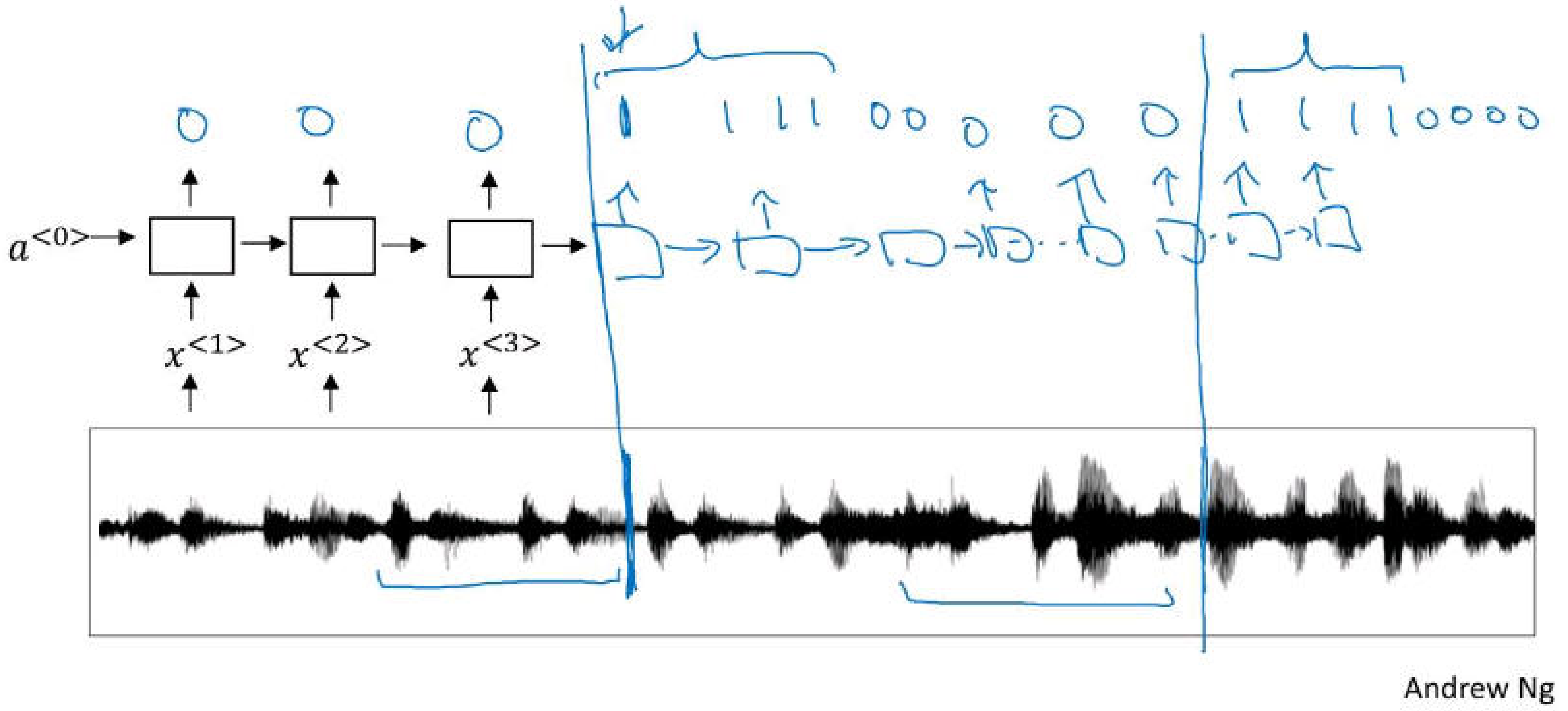
Basic rule: collapse repeated characters not separated by “blank”

[Graves et al., 2006. Connectionist Temporal Classification: Labeling unsegmented sequence data with recurrent neural networks] Andrew Ng

Trigger word detection algorithm

用0→1 标志

触发词检测



Andrew Ng